به نام خدا

تمرین دوم درس داده‌کاوی

ترم بهار ۱۴۰۲

توضیحات تمرین:

- پاسخ به این تمرین به صورت انفرادی می‌باشد.
- لطفا سوالات را به ترتیب پاسخ دهید.
- در صورت ابهام درباره‌ی تمرین با ایمیل درس با تدریسیاران در ارتباط باشید.

dm.spring1402@gmail.com

- مهلت ارسال تمرین تا **ساعت ۲۳:۵۹ دقیقه روز جمعه ۲۲ اردیبهشت** می‌باشد.
- تمرین شامل دو بخش **تئوری و عملی** می‌باشد.
- فایل‌های ارسالی شما باید یک فایل pdf گزارش (شامل جواب سوالات تئوری و سوالات بخش عملی)، و همچنین شامل کدهای شما باشد، که لطفا آنها را تحت یک فایل zip بارگزاری نمایید.
- فرمت فایل zip شما باید به شکل زیر باشد:

HW2-[student_number].zip (برای مثال HW2-9800000.zip)

فهرست

بخش تئوری

سوال اول

یکی از مباحثی که در درخت تصمیم مطرح میشود هرس درخت [1] برای جلوگیری از بیش برازش است. توضیح دهید چرا نمیتوان از مجموعه داده جدا برای هرس درخت استفاده کرد؟ منظور این است که داده‌هایی که برای هرس استفاده میشوند با مجموعه داده‌ای که برای ساخت درخت استفاده میشود یکسان نباشند.

سوال دوم

با توجه به مطالب تدریس شده در کلاس، برای داده‌های زیر یک درخت تصمیم درست کنید. (ذکر تمام مراحل و توضیح آنها لازم است)

آیا به مهمانی دعوت میشود؟	وزن	قد	رنگ لباس
خیر	لاغر	۱۷۰	قرمز
بله	چاق	۱۶۲	آبی
خیر	چاق	۱۶۵	سبز
بله	لاغر	۱۷۲	سبز
بله	لاغر	۱۶۰	آبی

سوال سوم

در جدول داده شده زیر با استفاده از قانون بیز برچسب داده زیر را به دست آورید. در صورت صفر شدن احتمال از هموارسازی لاپلاس [2] استفاده کنید.

(معدل = عالی ، مطالعه = بله ، حضور = خیر)

پاس شدن	حضور در کلاس‌ها	مطالعه برای امتحان	معدل
خیر	خیر	خیر	ضعیف
بله	بله	بله	ضعیف
خیر	خیر	خیر	متوسط
بله	بله	بله	متوسط
بله	خیر	خیر	عالی
بله	بله	بله	عالی

[1] Tree pruning
[2] laplace smoothing

سوال چهارم

همانطور که میدانیم یکی از معیارها برای ارزیابی مدل‌های یادگیری نظارت شده صحت [3] است. اما این معیار در برخی موارد ممکن است معیار مناسبی برای ارزیابی نباشد. موقعیت‌هایی که این معیار برای ارزیابی به خوبی عمل نمیکند را توضیح دهید.

سوال پنجم

فرض کنید که برای انتخاب پارامتر α در مدل از روش 10 fold cross validation استفاده کرده‌ایم. بهترین روش برای انتخاب مدل نهایی و تخمین ارور کدام است؟

سوال ششم

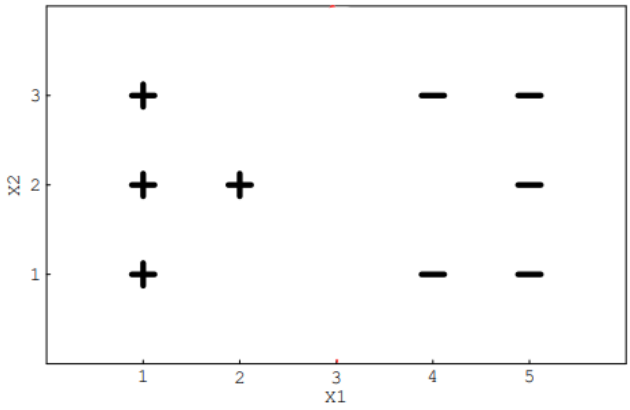
در الگوریتم boosting اگر هر کدام از موارد زیر رخ دهد ما یادگیری را متوقف میکنیم؟ برای پاسخ‌های خود دلیل بیاورید.

- میزان خطای طبقه‌بندی‌کننده ترکیبی در داده‌های آموزشی اصلی ۰ شود.
- میزان خطای طبقه‌بندی‌کننده ضعیف[4] فعلی روی داده‌های تمرین وزن‌دار[5] ۰ است.

سوال هفتم

فرض کنید برای داده‌های زیر از طبقه‌بندی کننده SVM خطی بدون کرنل استفاده میکنیم و پارامتر C در این طبقه بندی کننده بسیار بزرگ در نظر گرفته شده است. (اگر در مورد این پارامتر اطلاعی ندارید این لینک را مطالعه کنید).

الف) خطی که SVM گفته شده با استفاده از آن داده ها را دسته میکند را رسم کنید و علت انتخاب این خط را توضیح دهید.



ب) در شکل بالا نقاطی را انتخاب کنید که حذف آنها باعث میشود خطی که SVM داده ها را جدا میکند متفاوت از حالت (**الف**) شود. دلیل انتخاب این نقاط را توضیح دهید.

[3] Accuracy
[4] Weak
[5] Weighted training data

سوال هشتم

صحیح یا غلط بودن موارد زیر را با دلیل مشخص کنید :

الف) الگوریتم بیز ساده نمیتواند وابستگی بین متغیرها را مشخص کند.

ب) هنگامی که یک درخت تصمیم به سمت یک درخت پر پیش میرود احتمال اینکه نویز را هم پوشش دهد بیشتر میشود.

ج) در روش k نزدیک ترین همسایه[6] اگر k=1 الگوریتم نسبت به داده‌های نویز مقاوم تر از حالتی است که k=5 در نظر گرفته شود.

سوال نهم

فرض کنید در حال طراحی یک سیستم برای تشخیص خستگی راننده در اتومبیل هستید. بسیار مهم است که مدل شما خستگی را تشخیص دهد تا از هر گونه حادثه ای جلوگیری شود. کدام یک از معیارهای زیر بهترین معیار برای ارزیابی هست : Accuracy, Precision, Recall, Loss Value دلیل انتخاب خود را شرح دهید.

سوال دهم

علاوه بر شاخص آنتروپی برای ساخت درخت تصمیم، شاخص دیگری نیز وجود دارد که میتوان به جای آنتروپی از آن برای ساخت درخت استفاده کرد. این شاخص را معرفی کنید و بگویید تفاوت آن با آنتروپی چیست؟ بالا یا پایین بودن این شاخص چه معنایی دارد و چگونه محاسبه میشود.

سوال یازدهم

درمورد مسائل رگرسیون به سوالات زیر پاسخ دهید :

الف) simple linear regression و multiple linear regression با یکدیگر مقایسه کرده و تفاوت و شباهت های آنها را بیان کنید.

ب) یکی از راه‌های جلوگیری از بیش برازش استفاده از منظم‌سازی[7] است که به دو نوع L1 و L2 تقسیم میشود. به نوع اول Lasso Regression و به نوع دوم Ridge regression گفته میشود. تفاوت این دو روش را از نوع بهینه سازی بیان کرده و نحوه کار آنها را توضیح دهید.

ج) در جدول زیر سن و فشار خون چند بیمار قلبی داده شده است. معادله رگرسیون به فرم $y = \beta_0 + \beta_1 x$ به دست آورید. همچنین با استفاده از معادله به دست آمده فشار خون یک بیمار ۴۰ ساله را پیش بینی کنید.(متغیر x نشان دهنده سن و متغیر y نشان دهنده فشار خون است)

Patient	A	B	C	D	E	F	G
x	42	74	48	35	56	26	60
y	98	130	120	88	182	80	135

[6] K nearest neighbor
[7] regularization

بخش پیاده‌سازی

در این بخش شما باید با استفاده از الگوریتم‌های یادگیری ماشین و یادگیری عمیق که در دسته‌بندی داده‌ها استفاده می‌شوند، به پیش‌بینی پیش‌بینی قیمت خانه در تهران بپردازید.

برای پیاده‌سازی الگوریتم‌های یادگیری ماشین باید از کتابخانه scikit-learn و برای پیاده‌سازی شبکه عصبی باید از کتابخانه TensorFlow استفاده کنید. همچنین پیشنهاد می‌شود از Jupyter Notebook استفاده کنید و توضیحات لازم را در آن بنویسید.

مجموعه داده

مجموعه داده در نظر گرفته شده برای این تمرین مجموعه داده مربوط به پیش‌بینی قیمت خانه می‌باشد. این مجموعه شامل ویژگی‌های منطقه، تعداد اتاق، داشتن پارکینگ، انباری، آسانسور، آدرس و قیمت خانه متناظر با آن‌ها می‌شود.

مجموعه داده تحت یک فایل csv در اختیار شما قرار داده شده است. لازم است در ابتدا آن را دانلود کرده و با استفاده از کتابخانه Pandas بخوانید.

آماده‌سازی داده‌ها

پیش از آن که الگوریتم‌ها را روی داده‌ها پیاده‌سازی کنید نیاز است چند مرحله پیش‌پردازش روی داده‌ها انجام دهید.

- در این مجموعه داده تعدادی داده از دست رفته وجود دارد که باید سطر مربوط به مقادیر از دست رفته را حذف کنید.
- برچسب‌های این مجموعه داده (ستون قیمت) مقادیر پیوسته هستند. برای استفاده از الگوریتم‌های دسته‌بندی لازم است یک ستون به نام *priceLevel* به دیتافریم داده‌ها اضافه کنید که ستون قیمت را بر اساس چارک اول، دوم و سوم به چهار کلاس با برچسب‌های زیر تقسیم کند:

cheap: 0-25%
underMean: 25%-50%
upperMean: 50%-75%
expensive: 75%-100%

برای پیدا کردن چارک‌ها می‌توانید تابع `describe()` را روی ستون قیمت دیتافریم صدا بزنید.

- داده‌های categorical را به داده‌های عددی تبدیل کنید. برای اینکار می‌توانید از label encoding در کتابخانه scikit-learn استفاده کنید.
- روی فیچرها نرمال‌سازی انجام دهید. (دقت کنید که نرمال‌سازی را روی ستون *priceLevel* انجام ندهید.)
- داده‌ها را با نسبت ۸۰ به ۲۰ تقسیم[8] کنید. در واقع ۸۰ درصد داده‌ها به عنوان داده آموزشی برای آموزش مدل و ۲۰ درصد داده‌ها به عنوان داده تست استفاده شود.

[8] Splitting

پیش‌بینی با استفاده از الگوریتم‌های یادگیری ماشین

الگوریتم‌های زیر را با استفاده از کتابخانه scikit-learn روی داده‌های آموزشی آموزش دهید سپس با استفاده از داده‌های تست، ستون قیمت را پیش‌بینی کنید. برای تمامی الگوریتم‌ها دقت مدل را برای هم مجموعه آموزش و هم برای مجموعه تست گزارش کنید.

۱) رگرسیون[9]:

برای الگوریتم رگرسیون ستون *Price* را به عنوان برچسب در نظر بگیرید (ستون *priceLevel* را حذف کنید.) و بقیه ستون‌ها را به عنوان فیچر در نظر بگیرید.

- از رگرسیون خطی استفاده کنید.
- از رگرسیون چند جمله‌ای درجه ۲ و درجه ۳ استفاده کنید.
- در هر مورد دقت و خطای میانگین مربعات[10] را بدست آورید و نتایج را مقایسه کنید.

۲) دسته‌بندی[11]:

برای الگوریتم‌های دسته‌بندی ستون *priceLevel* را به عنوان برچسب در نظر بگیرید و ستون *Price* را حذف کنید.

- **درخت تصمیم:** مجموعه داده را با الگوریتم درخت تصمیم با شاخص انتروپی[12] آموزش دهید.
- **جنگل تصادفی:** از شاخص انتروپی استفاده کنید.
- **KNN:** الگوریتم KNN را به ازای سه پارامتر مختلف k آموزش دهید و دقت را گزارش کنید.
- **SVM:** از SVM هم در حالت خطی و هم غیر خطی استفاده کنید.

[9] Regression
[10] Mean Squared Error
[11] Classification
[12] Entropy

دسته‌بندی داده‌ها با استفاده از یادگیری عمیق

برای انجام این قسمت پیشنهاد می‌شود در ابتدا زمانی را در سایت TensorFlow playground بگذرانید و با تغییر پارامترها و بررسی انواع مسائل، شرایط مختلف را بررسی کنید.

در این بخش می‌خواهیم یک طبقه‌بندی چند کلاسه توسط شبکه عصبی انجام دهیم. برچسب داده‌ها را *priceLevel* در نظر بگیرید. یک شبکه عصبی به انتخاب خودتان ایجاد کرده و آن را روی داده‌های آموزشی آموزش دهید و با امتحان کردن مقادیر مختلف برای هایپرپارامترها به بهترین مقادیر ممکن برسید.

راهنمایی: برای طبقه‌بندی چند کلاسه از تابع هزینه "categorical_crossentropy" استفاده می‌شود. توجه کنید که برچسب‌ها باید به صورت one-hot encoding باشند. برای انکود کردن برچسب‌ها می‌توانید از تابع `to_categorical()` در TensorFlow استفاده کنید.

دقت مدل برای داده‌های آموزش و تست را گزارش کنید و ماتریس درهم‌ریختگی[۱۳] را رسم کنید. همچنین دلیل عملکرد مناسب و انتخاب هایپرپارامترها را توضیح دهید.

[13] Confusion matrix